BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 7: e112151
doi: 10.3897/biss.7.112151

OPEN ACCESS

Conference Abstract

On the Long Tails of Specimen Data

Arturo H. H. Ariño ‡, §

‡ University of Navarra, Institute of Biodiversity and Environment BIOMA, Pamplona, Spain
§ Biodiversity Data Analytics and Environmental Quality Group BEQ, Pamplona, Spain

Corresponding author: Arturo H. H. Ariño (artarip@unav.es)

Received: 04 Sep 2023 | Published: 07 Sep 2023

Citation: Ariño AHH (2023) On the Long Tails of Specimen Data. Biodiversity Information Science and Standards 7: e112151. https://doi.org/10.3897/biss.7.112151

Abstract

A recent article by K.R. Johnson and I.F.P. Owens in Science (Johnson and Owens 2023) suggested that the 73 main natural history museums around the world collectively hold over 1 billion records of accessioned "specimens" (taken as collection units), a result remarkably close to, but obtained through a completely different method from, research published a decade earlier by A.H. Ariño in Biodiversity Informatics (Ariño 2010). Both sets of approaches have benefitted from information available at the Global Biodiversity Information Facility (GBIF), which in the intervening years has grown by an order of magnitude, although mostly through observation-based occurrences rather than through accretion of specimen records in collections. When comparing the estimated size of collections and the amount of digital data from those collections, there is still a huge gap, as there was then. Digitization efforts have been progressing, but they are still far from reaching the goal of bringing information about all specimens into the digital domain.

While the larger institutions may doubtlessly have greater overall resources to try and make their data available than smaller institutions, how do they compare in terms of data mobilization and sharing? Not surprisingly, the distribution of the collection sizes shows a long tail of small institutions that, nonetheless, are also embarking on digitization efforts. Will this long tail of science actually manage to have all their biodiversity data available sooner than the larger institutions? It is becoming more widely recognized that data usability is predicated on data becoming findable, accessible, interoperable and reusable (FAIR, Wilkinson et al. 2016). What could be the consequences of having a data availability bias towards having many tiny collections available for ready use, rather than a much

© Ariño A. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

smaller (although surely very significant) fraction of larger collections of a comparable type?

This presentation explores and compares the distribution of potential versus readily available data in 2010 and in 2023, examines what trends might exist in the race to universal specimen data availability, and whether the digitization efforts might be better targeted to achieve greater overall scientific benefit.

Keywords

collections data, digitization, gaps

Presenting author

Arturo H. Ariño

Presented at

TDWG 2023

Acknowledgements

I am grateful to Stan Blum and Gail Kampmeier for their relevant comments and insights.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Ariño A (2010) Approaches to estimating the universe of natural history collections data. Biodiversity Informatics 7 (2). https://doi.org/10.17161/bi.v7i2.3991
- Johnson K, Owens IP (2023) A global approach for natural history museum collections. Science 379 (6638): 1192-1194. https://doi.org/10.1126/science.adf6434

- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18